

# A Constraint Satisfaction Method for Configuring Non-Local Service Interfaces

Pavel Zaichenkov, Olga Tveretina, and Alex Shafarenko

University of Hertfordshire, United Kingdom

**Abstract** Modularity and decontextualisation are core principles of a service-oriented architecture. However, the principles are often lost when it comes to an implementation of services, as a result of a rigidly defined service interface. The interface, which defines a data format, is typically specific to a particular context and its change entails significant redevelopment costs. This paper focuses on a two-fold problem. On the one hand, the interface description language must be flexible enough for maintaining service compatibility in a variety of different contexts without modification of the service itself. On the other hand, the composition of interfaces in a distributed environment must be provably consistent. The existing approaches for checking compatibility of service choreographies are either inflexible (WS-CDL and WSCI) or require behaviour specification associated with each service, which is often impossible to provide in practice.

We present a novel approach for automatic interface configuration in distributed stream-connected components operating as closed-source services (i.e. the behavioural protocol is unknown). We introduce a Message Definition Language (MDL), which can extend the existing interfaces description languages, such as WSDL, with support of subtyping, inheritance and polymorphism. The MDL supports configuration variables that link input and output interfaces of a service and propagate requirements over an application graph. We present an algorithm that solves the interface reconciliation problem using constraint satisfaction that relies on Boolean satisfiability as a subproblem.

## 1 Introduction

For the last decade service-oriented computing (SOC) has been a promising technology facilitating development of large scale distributed systems. SOC allows enterprises to expose their internal business systems as services available on the Internet. On the other hand, clients can combine services and reuse them for developing their own applications or constructing more complex services. Although web services continue to play an important role in modern software development, a service composition is still a key challenge for SOC and web services. Web service composition empowers organisations to build inter-enterprise software, to outsource software modules, and to provide an easily accessible functionality for their customers. Furthermore, service composition reduces the cost and risks of

new software development, because the software elements that are represented as web services can be reused repeatedly [19].

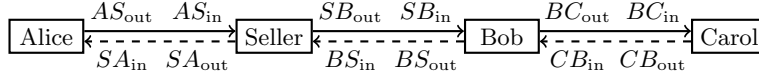
Web Service Description Language (WSDL) is an XML-based specification language for describing service interfaces, which is a *de facto* standard in SOC. Functionality and compatible data formats of the service are specified in WSDL in the form of an interface. The names and formats in the interfaces of communicating services must exactly match for interface compatibility. Today the environment in which services are developed and executed has become more open, changing and dynamic, which requires an adaptable and flexible approach to service composition. The choreography wired to specific WSDL interfaces is too restrictive for dependable service composition. The choreography is statically bounded to specific operation names and types, which impedes reusability of compound services and their interaction descriptions.

Reliable and dependable service composition remains a significant challenge today [19,10,18]. Services are provided autonomously by various organisations. Developers of applications, particularly safety-critical applications, such as health care, stock trading, nuclear systems, must be able to check soundness and completeness of service composition at early stages. Therefore, model checking and verification of web services is being actively researched today [18,5,21].

Web Services Choreography Description Language (WS-CDL) [13] and Web Service Choreography Interface (WSCI) [1] are languages for describing protocols from a global perspective. This approach is based on  $\pi$ -calculus that defines a behavioural semantics for concurrent processes. An application designer writes a global description in WS-CDL or WSCI that should be realisable by local protocols of communicating services. Service interfaces in WS-CDL are specified in WSDL. The relation between service interfaces connected with a communication channel is one-to-one, i.e. there is no way to propagate data format requirements and capabilities across the communication graph if services are not explicitly connected by a channel. Moreover, [2] emphasises that the existing association between WS-CDL and WSDL does not allow equivalent services with different WSDL interfaces to be part of the choreography.

Session types is another approach based on  $\pi$ -calculus that assures communication safety in distributed systems and in service choreographies particularly [6]. A choreography is defined as a global protocol in terms of the interactions that are expected from the protocol peers and a set of local protocols, one for each peer, which describes the global protocol from the viewpoint of an individual peer. The session types require services to expose their behaviour as a protocol. This information is enough to define a communication type system, which is well-suited for verifying runtime properties of the system such as deadlock-freedom, interleaving, etc. The session types essentially rely on behavioural protocols, which in most cases are neither explicitly provided nor can be derived from the code.

In this paper we present a formal method for configuring flexible interfaces based on constraint satisfaction and SAT. In contrast to the approaches based



**Figure 1.** Service composition in a Three Buyer usecase

on  $\pi$ -calculus, our method does not require services to define a protocol, but only to specify the data interface.

## 2 Motivating Example

Our approach for configuring web services is motivated by rapid development of Cloud computing, social networks and Internet of Things, which accelerate the growth and complexity of service choreographies [4,9,17]. Accordingly, we chose a simple but non-trivial example from one of those areas to illustrate our approach. The same example, known as the *three-buyer use case*, is often called upon to demonstrate the capabilities of session types such as communication safety, progress and session fidelity guarantees [7,14].

Consider a system involving buyers called Alice, Bob and Carol that cooperate in order to buy a book from a Seller. Each buyer is specified as an independent service that is connected with other services via a channel-based communication. There is an interface associated with every input and output port of a service, which specifies the service's functionality and data formats that the service is compatible with. The interfaces are defined in a Message Definition Language (MDL) that is formally introduced in Sect. 3. Fig. 1 depicts composition of the application where Alice is connected to Seller only and can interact with Bob and Carol indirectly.  $AS, SB, BC, CB, BS, AS$  denote interfaces that are associated with service input/output ports. For brevity, we only provide  $AS, SB$  and  $BC$  (the rest of the interfaces are defined in the same manner), which are specified in the MDL as terms in the following way:

$$\begin{array}{l|l}
 AS_{out} = (:request: \{title: tv^\downarrow\}, & AS_{in} = (:request: \{title: string\}, \\
 \quad payment: \{title: tv^\downarrow, money: int, id: int\}, & \quad payment: \{title: string, money: int\} \\
 \quad share(x): \{title: tv^\downarrow, money: int\}, & \quad | ct1^\uparrow:) \\
 \quad suggest(y): \{title: tv^\downarrow\}:) & SB_{in} = (:share(z): \{quote: string, \\
 SB_{out} = (:response: \{title: string, money: int\} & \quad money: int\}, \\
 \quad | ct1^\uparrow:) & \quad response: \{title: string, money: int\} \\
 BC_{out} = (:share(z): \{quote: string, money: int\} | ct2^\uparrow:) & BC_{in} = (:share: \{quote: string, money: int\}:)
 \end{array}$$

$(: :)$  delimit a collection of alternative label-record pairs called *variants*, where the label corresponds to the particular implementation that can process a message defined by the given record. A *record* delimited by  $\{ \}$  is a collection of label-value pairs. Collection elements may contain Boolean variables called *guards* (e.g.  $x$ ,  $y$  or  $z$  in our example). A guard instantiated to **false** excludes the element from the collection. This is the main self-configuration mechanism: Boolean variables control the dependencies between any elements of interface

collections (this can be seen as a generalised version of intersection types [8]) The variables exclude elements from the collection if the dependencies between corresponding elements in the interfaces that are connected by a communication channel cannot be satisfied.

Parametric polymorphism is supported using interface variables, such as  $tv^\downarrow$ ,  $ct1^\uparrow$  and  $ct2^\uparrow$  (the meaning of  $\uparrow$  and  $\downarrow$  is explain in Sect. 3). Moreover, the presence of  $ct1^\uparrow$  and  $ct2^\uparrow$  in both input and output interfaces enables flow inheritance [12] mechanism that provides delegation of the data and service functionality across available services.

$AS_{out}$  declares an output interface of Alice, which declares functionality and a format of messages sent to Seller. The service has the following functionality:

- Alice can **request** a book’s price from Seller by providing a **title** of an arbitrary type (which is specified by a term variable  $tv^\downarrow$ ) that Seller is compatible with. On the other hand, Seller declares that a title of type **string** is only acceptable, which means that  $tv^\downarrow$  must be instantiated to **string**.
- Furthermore, Alice can provide a **payment** for a book. In addition to the **title** and the required amount of **money**, Alice provides her **id** in the message. Although Seller does not require the **id**, the interconnection is still valid (a description in standard WSDL interfaces would cause an error though) due to the subtyping supported in the MDL.
- Furthermore, Alice can offer to **share** a purchase between other customers. Although Alice is not connected to Bob or Carol and may even not be aware of their presence (the example illustrates a composition where some service communicates with services that the service is not directly connected with), our mechanism detects that Alice can send a message with “**share**” label to Bob by bypassing it implicitly through Seller. In order to enable inheritance in Seller’s service, the mechanism sets a tail variable  $ct1^\uparrow$  to  $(:share: \{title: string, money: int\} :)$ . If Bob were unable to accept a message with “**share**” label, the mechanism would instantiate  $x$  with **false**, which automatically removes the corresponding functionality from the service.
- Finally, Alice can **suggest** a book to other buyers. However, examination of other service interfaces shows that there is no service that can receive a message with the label “**suggest**”. Therefore, a communication error occurs if Alice decides to send the message. To avoid this, the configuration mechanism excludes “**suggest**” functionality from Alice’s service by setting  $y$  variable to **false**.

The proposed configuration mechanism analyses the interfaces of services Seller, Bob and Carol in the same manner. The presence of  $ct1^\uparrow$  variable in both input and output interfaces of Bob enables support of data inheritance on the interface level. Furthermore, the Boolean variable  $z$  behaves as an intersection type: Bob has “purchase sharing” functionality declared as an element  $share(z): \{...\}$  in its input interface  $SB_{in}$  (used by Seller). The element is related to the element  $share(z): \{...\}$  in its output interface  $BC_{out}$  (used by Carol). The relation declares that Bob provides Carol with “sharing” functionality only

if Bob was provided with the same functionality from Seller. In our example,  $z$  is true, because Carol declares that it can receive messages with the label “share”. Note that there could be an any Boolean formula in place of  $z$ , which wires any input and output interfaces of a single service in an arbitrary way. The existing interface description languages (WSDL, WS-CDL, etc.) do not support such interface wiring capabilities.

Interface variables provide facilities similar to C++ templates. Services can specify a generic behaviour compatible with multiple contexts and input/output data formats. Given the context, the compiler then specialises the interfaces based on the requirements and capabilities of other services.

The problem being solved is similar to type inference problem; however, it has large combinatorial complexity and, therefore, direct search of a solution is impractical. Furthermore, additional complexity arises from the presence of Boolean variables in general form. Another problem is potential cyclic dependencies in the network, which prevent the application of a simple forward algorithm. In our approach, we define our problem as a constraint satisfaction problem. Then we employ a constraint solver, which was specifically developed to solve this problem, to find correct instantiations of the variables.

### 3 Message Definition Language and CSP

Now we define a term algebra called Message Definition Language (MDL). The purpose of the MDL is to describe flexible service interfaces. Although we use a concise syntax for MDL terms that is different from what standard WSDL-based interfaces look like, it can easily be rewritten as a WSDL extension.

In our approach, a message is a collection of data entities, each specified by a corresponding *term*. The intention of the term is to represent

1. a standard atomic type such as `int`, `string`, etc.;
2. inextensible data collections such as tuples;
3. extensible data records [11,16], where additional named fields can be introduced without breaking the match between the producer and the consumer and where fields can also be inherited from input to output records by lowering the output type, which is always safe;
4. data-record variants, where generally more variants can be accepted by the consumer than the producer is aware of, and where such additional variants can be inherited from the output back to the input of the producer — hence contravariance — again, by raising the input type, which is always safe, too.

#### 3.1 Terms

Each term is either atomic or a collection in its own right. Atomic terms are *symbols*, which are identifiers used to represent standard types such as `int`, `string`, etc. To account for subtyping we include three categories of collections: *tuples* that are demanded to be of the same size and thus admit only depth

structural subtyping, *records* that are subtyped covariantly (a larger record is a subtype) and *choices* that are subtyped contravariantly using set inclusion (a smaller choice is a subtype).

In order to support parametric polymorphism and inheritance in interfaces, we introduce term variables (called later *t-variables*), which are similar to type variables. For coercion of interfaces it is important to distinguish between two variable categories: *down-coerced* and *up-coerced* ones. The former can be instantiated with symbols, tuples and records (terms of these three categories are called down-coerced terms), and the latter can only be instantiated with choices (up-coerced terms). Informally, for two down-coerced terms, a term associated with a structure with “more data” is a subtype of the one associated with a structure that contains less; and vice versa for up-coerced terms. We use the notation  $v^\downarrow$  and  $v^\uparrow$  for down-coerced and up-coerced variables respectively, and  $v$  when its coercion sort is unimportant. Explicit sort annotation on variables is useful for simplifying partial order definitions on terms.

We introduce Boolean variables (called *b-variables* below) in the term interfaces to specify dependencies between input and output data formats. B-variables provide functionality similar to intersection types, which increase the expressiveness of function signatures.

A Boolean expression  $b \in \mathcal{B}$  ( $\mathcal{B}$  denotes a set of Boolean expressions) called a guard is defined by the following grammar:

$$\begin{aligned} \langle \text{guard} \rangle \quad &::= (\langle \text{guard} \rangle \wedge \langle \text{guard} \rangle) \mid (\langle \text{guard} \rangle \vee \langle \text{guard} \rangle) \mid \langle \text{guard} \rangle \rightarrow \langle \text{guard} \rangle \mid \\ &\neg \langle \text{guard} \rangle \mid \mathbf{true} \mid \mathbf{false} \mid b\text{-variable} \end{aligned}$$

MDL terms are built recursively using the constructors: tuple, record, choice and switch, according to the following grammar:

$$\begin{aligned} \langle \text{term} \rangle \quad &::= \langle \text{symbol} \rangle \mid \langle \text{tuple} \rangle \mid \langle \text{record} \rangle \mid \langle \text{choice} \rangle \mid \text{t-variable} \\ \langle \text{tuple} \rangle \quad &::= (\langle \text{term} \rangle \mid \langle \text{term} \rangle^*) \\ \langle \text{record} \rangle \quad &::= \{[\langle \text{element} \rangle], \langle \text{element} \rangle^* \mid \text{down-coerced t-variable}\} \\ \langle \text{choice} \rangle \quad &::= (:\langle \text{element} \rangle], \langle \text{element} \rangle^* \mid \text{up-coerced t-variable}): \\ \langle \text{element} \rangle \quad &::= \langle \text{label} \rangle \langle \langle \text{guard} \rangle \rangle : \langle \text{term} \rangle \\ \langle \text{label} \rangle \quad &::= \langle \text{symbol} \rangle \end{aligned}$$

Informally, a *tuple* is an ordered collection of terms and a *record* is an extensible, unordered collection of guarded labeled terms, where *labels* are arbitrary symbols, which are unique within a single record. A *choice* is a collection of alternative terms. The syntax of choice is the same as that of record except for the delimiters. The difference between records and choices is in width subtyping and will become clear below when we define seniority on terms. We use choices to represent polymorphic messages and service interfaces on the top level. Records and choices are defined in *tail form*. The tail is denoted by a t-variable that represents a term of the same kind as the construct in which it occurs.

A switch is an auxiliary construct intended for building conditional terms, which is specified as a set of unlabelled (by contrast to a choice) guarded alternatives. Formally, it is defined as

$\langle \text{switch} \rangle ::= \langle \langle \text{guard} \rangle : \langle \text{term} \rangle [, \langle \text{guard} \rangle : \langle \text{term} \rangle]^* \rangle$

Exactly one guard must be **true** for any valid switch, i.e. the switch is substitutionally equivalent to the term marked by the **true** guard:

$$\langle (\text{false}) : t_1, \dots, (\text{true}) : t_i, \dots, (\text{false}) : t_n \rangle = \langle (\text{true}) : t_i \rangle = t_i.$$

For example,  $\langle (a) : \text{int}, (\neg a) : \text{string} \rangle$  represents the symbol *int* if  $a = \text{true}$ , and the symbol *string* otherwise.

### 3.2 Seniority Relation

For a guard  $g$ , we denote as  $V^b(g)$  the set of b-variables that occur in  $g$ . For a term  $t$ , we denote as  $V^\uparrow(t)$  the set of up-coerced t-variables that occur in  $t$ , and as  $V^\downarrow(t)$  the set of down-coerced ones; and finally  $V^b(t)$  is the set of b-variables in  $t$ .

**Definition 1 (Semi-ground and ground terms).** *A term  $t$  is called semi-ground if  $V^\uparrow(t) \cup V^\downarrow(t) = \emptyset$ . A term  $t$  is called ground if it is semi-ground and  $V^b(t) = \emptyset$ .*

**Definition 2 (Well-formed terms).** *A term  $t$  is well-formed if it is ground and exactly one of the following holds:*

1.  $t$  is a symbol;
2.  $t$  is a tuple  $(t_1 \dots t_n)$ ,  $n > 0$ , where all  $t_i$ ,  $1 \leq i \leq n$ , are well-formed;
3.  $t$  is a record  $\{l_1(g_1) : t_1, \dots, l_n(g_n) : t_n\}$  or a choice  $(:l_1(g_1) : t_1, \dots, l_n(g_n) : t_n:)$ ,  $n \geq 0$ , where for all  $1 \leq i \neq j \leq n$ ,  $g_i \wedge g_j \implies l_i \neq l_j$  and all  $t_i$  for which  $g_i$  are **true** are well-formed;
4.  $t$  is a switch  $\langle (g_1) : t_1, \dots, (g_n) : t_n \rangle$ ,  $n > 0$ , where for some  $1 \leq i \leq n$ ,  $g_i = \text{true}$  and  $t_i$  is well-formed and where  $g_j = \text{false}$  for all  $j \neq i$ .

If an element of a record, choice or switch has a guard that is equal to **false**, then the element can be omitted, e.g.

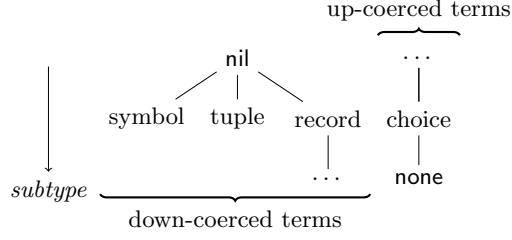
$$\{\mathbf{a}(x \wedge y) : \text{string}, \mathbf{b}(\text{false}) : \text{int}, \mathbf{c}(x) : \text{int}\} = \{\mathbf{a}(x \wedge y) : \text{string}, \mathbf{c}(x) : \text{int}\}.$$

If an element of a record or a choice has a guard that is **true**, the guard can be syntactically omitted, e.g.

$$\{\mathbf{a}(x \wedge y) : \text{string}, \mathbf{b}(\text{true}) : \text{int}, \mathbf{c}(x) : \text{int}\} = \{\mathbf{a}(x \wedge y) : \text{string}, \mathbf{b} : \text{int}, \mathbf{c}(x) : \text{int}\}.$$

We define the *canonical form* of a well-formed collection as a representation that does not include **false** guards, and we omit **true** guards anyway. The canonical form of a switch is its (only) term with a **true** guard, hence any term in canonical form is switch-free.

Next we introduce a seniority relation on terms for the purpose of structural subtyping. In the sequel we use **nil** to denote the empty record  $\{ \}$ , which has the meaning of unit type and represents a message without any data. Similarly, we use **none** to denote the empty choice  $(: :)$ .



**Figure 2.** Two semilattices representing the seniority relation for terms of different categories. The lower terms are the subtypes of the upper ones

**Definition 3 (Seniority relation).** *The seniority relation  $\sqsubseteq$  on well-formed terms is defined in canonical form as follows:*

1.  $\text{none} \sqsubseteq t$  if  $t$  is a choice;
2.  $t \sqsubseteq \text{nil}$  if  $t$  is a symbol, a tuple or a record;
3.  $t \sqsubseteq t$ ;
4.  $t_1 \sqsubseteq t_2$ , if for some  $k, m > 0$  one of the following holds:
  - (a)  $t_1 = (t_1^1 \dots t_1^k)$ ,  $t_2 = (t_2^1 \dots t_2^k)$  and  $t_1^i \sqsubseteq t_2^i$  for each  $1 \leq i \leq k$ ;
  - (b)  $t_1 = \{l_1^1: t_1^1, \dots, l_1^k: t_1^k\}$  and  $t_2 = \{l_2^1: t_2^1, \dots, l_2^m: t_2^m\}$ , where  $k \geq m$  and for each  $j \leq m$  there is  $i \leq k$  such that  $l_1^i = l_2^j$  and  $t_1^i \sqsubseteq t_2^j$ ;
  - (c)  $t_1 = (:l_1^1: t_1^1, \dots, l_1^k: t_1^k)$  and  $t_2 = (:l_2^1: t_2^1, \dots, l_2^m: t_2^m)$ , where  $k \leq m$  and for each  $i \leq k$  there is  $j \leq m$  such that  $l_1^i = l_2^j$  and  $t_1^i \sqsubseteq t_2^j$ .

Similarly to the t-variables, terms are classified into two categories: symbols, tuples and records are down-coerced terms and choices are up-coerced terms. The seniority relation defines a symmetric relation on down-coerced and up-coerced terms: an element  $\text{nil}$  is the maximum element for down-coerced terms; on the other hand,  $\text{none}$  is the minimum element for up-coerced terms.  $\mathcal{T}^\downarrow$  denotes the set of all down-coerced ground terms,  $\mathcal{T}^\uparrow$  denotes the set of all up-coerced ground terms and  $\mathcal{T} = \mathcal{T}^\downarrow \cup \mathcal{T}^\uparrow$  is the set of all ground terms. Similarly,  $\mathcal{T}_m^\downarrow$  denotes the set of all vectors of down-coerced ground terms of length  $m$  and  $\mathcal{T}_n^\uparrow$  denotes the set of all vectors of up-coerced ground terms of length  $n$ . If  $\mathbf{t}_1$  and  $\mathbf{t}_2$  are vectors of terms  $(t_1^1, \dots, t_n^1)$  and  $(t_2^1, \dots, t_n^2)$  of size  $n$ , then  $\mathbf{t}_1 \sqsubseteq \mathbf{t}_2$  denotes the seniority relation for all pairs  $t_i^1 \sqsubseteq t_i^2$  ( $1 \leq i \leq n$ ).

**Proposition 1.** *The seniority relation  $\sqsubseteq$  is a partial order, and  $(\mathcal{T}, \sqsubseteq)$  is a pair of meet and join semilattices (Fig. 2):*

$$\begin{aligned} \forall t_1, t_2 \in \mathcal{T}^\downarrow, t_1 \sqsubseteq t_2 \text{ iff } t_1 \sqcap t_2 = t_1; \\ \forall t_1, t_2 \in \mathcal{T}^\uparrow, t_1 \sqsubseteq t_2 \text{ iff } t_1 \sqcup t_2 = t_2. \end{aligned}$$

The seniority relation represents the subtyping relation on terms. If a term  $t$  describes the input interface of a service, then the service can process any message described by a term  $t'$ , such that  $t' \sqsubseteq t$ .



Although the seniority relation is straightforwardly defined for ground terms, terms that are present in the interfaces of services can contain t-variables and b-variables. Finding such ground term values for the t-variables and such Boolean values for the b-variables that the seniority relation holds represents a CSP problem, which is formally introduced next.

### 3.3 Constraint Satisfaction Problem for Web Services

We define a substitution, which is used in the definition of the CSP and in the algorithm, as a syntactic transformation that replaces b-variables with Boolean values and t-variables with ground or semi-ground values.

**Definition 4 (Substitution).** Let  $g$  be a guard,  $t$  be a term,  $k = |\mathbf{V}^b(g) \cup \mathbf{V}^b(t)|$ , and  $\mathbf{f} = (f_1, \dots, f_k)$  be a vector of b-variables contained in  $g$  and  $t$ , and  $\mathbf{v} = (v_1, \dots, v_k)$  be a vector of term variables contained in  $t$ . Then for any vector of Boolean values  $\mathbf{b} = (b_1, \dots, b_k)$  and a vector of terms  $\mathbf{s} = (s_1, \dots, s_k)$

1.  $g[\mathbf{f}/\mathbf{b}]$  denotes a Boolean value (true or false), which is obtained as a result of the simultaneous replacement and evaluation of  $f_i$  with  $b_i$  for each  $1 \leq i \leq k$ ;
2.  $t[\mathbf{f}/\mathbf{b}]$  denotes the vector obtained as a result of the simultaneous replacement of  $f_i$  with  $b_i$  for each  $1 \leq i \leq k$ ;
3.  $t[\mathbf{v}/\mathbf{s}]$  denotes the vector obtained as a result of the simultaneous replacement of  $v_i$  with  $s_i$  for each  $1 \leq i \leq k$ ;
4.  $t[\mathbf{f}/\mathbf{b}, \mathbf{v}/\mathbf{s}]$  is a shortcut for  $t[\mathbf{f}/\mathbf{b}][\mathbf{v}/\mathbf{s}]$ .

Given the set of constraints  $\mathcal{C}$ , we define the set of b-variables as

$$\mathbf{V}^b(\mathcal{C}) = \bigcup_{t \sqsubseteq t' \in \mathcal{C}} \mathbf{V}^b(t) \cup \mathbf{V}^b(t'),$$

the sets of of down-coerced and up-coerced t-variables as

$$\mathbf{V}^\downarrow(\mathcal{C}) = \bigcup_{t \sqsubseteq t' \in \mathcal{C}} \mathbf{V}^\downarrow(t) \cup \mathbf{V}^\downarrow(t') \quad \text{and} \quad \mathbf{V}^\uparrow(\mathcal{C}) = \bigcup_{t \sqsubseteq t' \in \mathcal{C}} \mathbf{V}^\uparrow(t) \cup \mathbf{V}^\uparrow(t').$$

In the following for each set of constraints  $S$  such that  $|\mathbf{V}^b(S)| = l$ ,  $|\mathbf{V}^\uparrow(S)| = m$  and  $|\mathbf{V}^\downarrow(S)| = n$  we use  $\mathbf{f} = (f_1, \dots, f_l)$  to denote the vector of b-variables contained in  $S$ ,  $\mathbf{v}^\uparrow = (v_1^\uparrow, \dots, v_m^\uparrow)$  to denote the vector of up-coerced t-variables and  $\mathbf{v}^\downarrow = (v_1^\downarrow, \dots, v_n^\downarrow)$  to denote the vector of down-coerced t-variables.

Let  $\mathcal{C}$  be a set of constraints such that  $|\mathbf{V}^b(\mathcal{C})| = l$ ,  $|\mathbf{V}^\downarrow(\mathcal{C})| = m$ ,  $|\mathbf{V}^\uparrow(\mathcal{C})| = n$  and for some  $l, m, n \geq 0$ . Now we can define a CSP-WS formally as follows.

**Definition 5 (CSP-WS).** Find a vector of Boolean values  $\mathbf{b} = (b_1, \dots, b_l)$  and vectors of ground terms  $\mathbf{t}^\downarrow = (t_1^\downarrow, \dots, t_m^\downarrow)$ ,  $\mathbf{t}^\uparrow = (t_1^\uparrow, \dots, t_n^\uparrow)$ , such that for each  $t_1 \sqsubseteq t_2 \in \mathcal{C}$

$$t_1[\mathbf{f}/\mathbf{b}, \mathbf{v}^\downarrow/\mathbf{t}^\downarrow, \mathbf{v}^\uparrow/\mathbf{t}^\uparrow] \sqsubseteq t_2[\mathbf{f}/\mathbf{b}, \mathbf{v}^\downarrow/\mathbf{t}^\downarrow, \mathbf{v}^\uparrow/\mathbf{t}^\uparrow]$$

The tuple  $(\mathbf{b}, \mathbf{t}^\downarrow, \mathbf{t}^\uparrow)$  is called a solution.

## 4 Solution Approximation

One way to solve CSP-WS is to attempt to solve the problem for all possible instantiations of b-variables. We start with considering the simplification when the original problem is reduced to the one without b-variables provided that some vector of Boolean assignments is given.

We use an approximation algorithm that iteratively traverses the meet and the join semilattices<sup>1</sup> for vectors of ground terms  $\mathcal{T}_m^\downarrow$  and  $\mathcal{T}_n^\uparrow$ , where  $m = |\mathcal{V}^\downarrow(\mathcal{C})|$  and  $n = |\mathcal{V}^\uparrow(\mathcal{C})|$ , which represent solution approximations for down-coerced and up-coerced terms respectively. The algorithm monotonically converges to a solution if one exists. Informally, the algorithm performs the following steps:

1. Compute the initial approximation of the solution for  $i = 0$  as  $(\mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow) = ((\text{nil}, \dots, \text{nil}), (\text{none}, \dots, \text{none}))$ , where the first element in the pair is the vector of top elements from the meet semilattice and the second element is the vector of bottom elements from the join semilattice.
2. Compute  $(\mathbf{a}_{i+1}^\downarrow, \mathbf{a}_{i+1}^\uparrow)$  such that  $\mathbf{a}_{i+1}^\downarrow \sqsubseteq \mathbf{a}_i^\downarrow$  and  $\mathbf{a}_i^\uparrow \sqsubseteq \mathbf{a}_{i+1}^\uparrow$ .
3. Repeat step 2 until a chain of approximations converges to the solution, i.e.  $(\mathbf{a}_{i+1}^\downarrow, \mathbf{a}_{i+1}^\uparrow) = (\mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow)$ , or a situation where some of the constraints from step 2 cannot be satisfied. Then return the last approximation as the solution or **Unsat**.

We extend the set  $\mathcal{T}_m^\downarrow$  with the element  $\underline{\perp}$ , i.e.  $\tilde{\mathcal{T}}_m^\downarrow = \mathcal{T}_m^\downarrow \cup \{\underline{\perp}\}$ , and the set  $\mathcal{T}_n^\uparrow$  with the element  $\overline{\top}$ , i.e.  $\tilde{\mathcal{T}}_n^\uparrow = \mathcal{T}_n^\uparrow \cup \{\overline{\top}\}$ . Here  $\underline{\perp}$  is defined as the bottom element of the meet semilattice, i.e.  $\underline{\perp} \sqsubseteq \mathbf{a}^\downarrow$  for any  $\mathbf{a}^\downarrow \in \tilde{\mathcal{T}}_m^\downarrow$ , and  $\overline{\top}$  is defined as the top element of the join semilattice, i.e.  $\mathbf{a}^\uparrow \sqsubseteq \overline{\top}$  for any  $\mathbf{a}^\uparrow \in \tilde{\mathcal{T}}_n^\uparrow$ . The algorithm returns  $\underline{\perp}$  or  $\overline{\top}$  if it is unable to find an approximation for some constraints, which, as shown in Theorem 2 below, means that the input set of constraints does not have a solution.

### 4.1 Approximating Function

In order to specify how the next approximation is computed we introduce the *approximating function*  $\text{AF} : \mathcal{C} \times \tilde{\mathcal{T}}_m^\downarrow \times \tilde{\mathcal{T}}_n^\uparrow \rightarrow \tilde{\mathcal{T}}_m^\downarrow \times \tilde{\mathcal{T}}_n^\uparrow$  that maps a single constraint and the current approximation to the new approximation.

The function  $\text{AF}$  is given below for all categories of terms (except for choices because they are symmetrical to the cases for records and switches that are reduced to other term categories). Let  $\mathbf{v}^\downarrow = (\bar{v}_1, \dots, \bar{v}_m)$ ,  $\mathbf{v}^\uparrow = (\bar{\bar{v}}_1, \dots, \bar{\bar{v}}_n)$ ,  $\mathbf{a}^\downarrow = (\bar{a}_1, \dots, \bar{a}_m)$ ,  $\mathbf{a}^\uparrow = (\bar{\bar{a}}_1, \dots, \bar{\bar{a}}_n)$ .

If  $t$  is a symbol, the given approximation  $(\mathbf{a}^\downarrow, \mathbf{a}^\uparrow)$  already satisfies the constraint:

$$\text{AF}(t \sqsubseteq t, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = (\mathbf{a}^\downarrow, \mathbf{a}^\uparrow).$$

If  $t$  is a down-coerced term and  $\bar{v}_l$  is a down-coerced variable, the approximation for  $\bar{v}_l$  is used to refine the approximation for variables in  $t$ . Therefore, the

<sup>1</sup> The algorithm presented here is widely used for data-flow analysis and flow graph optimisations [15].

constraint is reduced to the one with  $\bar{v}_l$  as a ground term, which is obtained by substitution  $\bar{v}_l[v^\downarrow/a^\downarrow]$ :

$$\text{AF}(t \sqsubseteq \bar{v}_l, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = \text{AF}(t \sqsubseteq \bar{v}_l[v^\downarrow/a^\downarrow], \mathbf{a}^\downarrow, \mathbf{a}^\uparrow).$$

If  $\bar{v}_l$  is an up-coerced variable and  $t$  is an up-coerced term, the case is symmetric to the previous one:

$$\text{AF}(\bar{v}_l \sqsubseteq t, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = \text{AF}(\bar{v}_l[v^\uparrow/a^\uparrow] \sqsubseteq t, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow).$$

If  $\bar{v}_l$  is a down-coerced variable and  $t$  is a down-coerced term, then  $\bar{v}_l$  must be not higher than the ground term  $t[v^\downarrow/a^\downarrow, v^\uparrow/a^\uparrow]$  in the meet semilattice:

$$\text{AF}(\bar{v}_l \sqsubseteq t, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = ((\bar{a}_1, \dots, \bar{a}_l \sqcap t[v^\downarrow/a^\downarrow, v^\uparrow/a^\uparrow], \dots, \bar{a}_m), \mathbf{a}^\uparrow).$$

If  $t$  is an up-coerced term and  $v_l^\uparrow$  is an up-coerced variable, the case is symmetric to the previous one:

$$\text{AF}(t \sqsubseteq v_l^\uparrow, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = (\mathbf{a}^\downarrow, (\bar{a}_1, \dots, \bar{a}_l \sqcup t[v^\downarrow/a^\downarrow, v^\uparrow/a^\uparrow], \dots, \bar{a}_n)).$$

If  $t_1$  and  $t_2$  are tuples  $(t_1^1 \dots t_k^1)$  and  $(t_1^2 \dots t_k^2)$  respectively, then the constraint must hold for the corresponding nested terms:

$$\text{AF}((t_1^1 \dots t_k^1) \sqsubseteq (t_1^2 \dots t_k^2), \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = (\prod_{1 \leq i \leq k} \mathbf{a}_i^\downarrow, \bigsqcup_{1 \leq i \leq k} \mathbf{a}_i^\uparrow).$$

If  $t_1$  and  $t_2$  are records  $\{l_1^1: t_1^1, \dots, l_p^1: t_p^1\}$  and  $\{l_1^2: t_1^2, \dots, l_q^2: t_q^2\}$  respectively, two cases must be considered:

- If for all  $i$  ( $1 \leq i \leq q$ ) there exists  $j$  such that  $l_j^1 = l_i^2$ , then the constraint for nested terms  $t_j^1 \sqsubseteq t_i^2$  must hold:

$$\text{AF}(\{l_1^1: t_1^1, \dots, l_p^1: t_p^1\} \sqsubseteq \{l_1^2: t_1^2, \dots, l_q^2: t_q^2\}, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = (\prod_{1 \leq i \leq q} \mathbf{a}_i^\downarrow, \bigsqcup_{1 \leq i \leq q} \mathbf{a}_i^\uparrow).$$

- Otherwise, the set of labels in  $t_2$  is not a subset of the labels in  $t_1$  and, therefore,  $t_1 \sqsubseteq t_2$  is unsatisfiable:

$$\text{AF}(\{l_1^1: t_1^1, \dots, l_p^1: t_p^1\} \sqsubseteq \{l_1^2: t_1^2, \dots, l_q^2: t_q^2\}, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = (\perp, \overline{\top}).$$

If  $\bar{v}_l$  is a down-coerced variable,  $t_1$  and  $t_2$  are records  $\{l_1^1: t_1^1, \dots, l_p^1: t_p^1 | \bar{v}_l\}$  and  $\{l_1^2: t_1^2, \dots, l_q^2: t_q^2\}$  respectively, the constraint can be satisfied only if for every nested term  $t_i^2$  with the label  $l_i^2$  in  $t$  one of the following holds: 1) there exists a subterm  $t_j^1$  with equal label in  $t_1$  and  $t_j^1 \sqsubseteq t_i^2$  holds, or 2)  $\bar{v}_l$  is a record that contains a junior to  $t_i^2$  element with the same label:

$$\text{AF}(\{l_1^1: t_1^1, \dots, l_p^1: t_p^1 | \bar{v}_l\} \sqsubseteq \{l_1^2: t_1^2, \dots, l_q^2: t_q^2\}, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = (\prod_{1 \leq i \leq q} \mathbf{a}_i^\downarrow, \bigsqcup_{1 \leq i \leq q} \mathbf{a}_i^\uparrow),$$

where

$$(\mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow) = \begin{cases} \text{AF}(t_j^1 \sqsubseteq t_i^2, \mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow) & \text{if } \exists j : l_j^1 = l_i^2 \\ ((\bar{a}_1, \dots, \bar{a}_l \sqcap t_i^2[\mathbf{v}^\downarrow/\mathbf{a}^\downarrow, \mathbf{v}^\uparrow/\mathbf{a}^\uparrow], \dots, \bar{a}_m), \mathbf{a}^\uparrow) & \text{otherwise.} \end{cases}$$

If  $t_1$  is a record  $\{l_1^1: t_1^1, \dots, l_p^1: t_p^1\}$  or  $\{l_1^1: t_1^1, \dots, l_p^1: t_p^1 \mid \bar{v}_l\}$  and  $t_2$  is a record  $\{l_1^2: t_1^2, \dots, l_q^2: t_q^2 \mid \bar{u}_r\}$ , then the constraint can by substitution be reduced to the previous cases for records:

$$\text{AF}(t_1 \sqsubseteq t_2, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = \text{AF}(t_1 \sqsubseteq t_2[\bar{u}_r/\bar{a}_r], \mathbf{a}^\downarrow, \mathbf{a}^\uparrow).$$

The function  $\text{AF}$  has the homomorphism property, which is important for showing termination and correctness of the algorithm.

**Lemma 1 (Homomorphism).** *Let  $\text{AF}(t_1 \sqsubseteq t_2, \mathbf{a}_1^\downarrow, \mathbf{a}_1^\uparrow) = (\bar{\mathbf{a}}_1^\downarrow, \bar{\mathbf{a}}_1^\uparrow)$  and  $\text{AF}(t_1 \sqsubseteq t_2, \mathbf{a}_2^\downarrow, \mathbf{a}_2^\uparrow) = (\bar{\mathbf{a}}_2^\downarrow, \bar{\mathbf{a}}_2^\uparrow)$ . Then*

$$\text{AF}(t_1 \sqsubseteq t_2, \mathbf{a}_1^\downarrow \sqcap \mathbf{a}_2^\downarrow, \mathbf{a}_1^\uparrow \sqcup \mathbf{a}_2^\uparrow) = (\bar{\mathbf{a}}_1^\downarrow \sqcap \bar{\mathbf{a}}_2^\downarrow, \bar{\mathbf{a}}_1^\uparrow \sqcup \bar{\mathbf{a}}_2^\uparrow).$$

The function  $\text{AF}_\mathcal{C}$  is a composition of  $\text{AF}$  functions that are sequentially applied to all constraints in  $\mathcal{C}$  (the order in which  $\text{AF}$  is applied to the constraints is not important due to distributivity of the semi-lattices):

$$\text{AF}_\mathcal{C}(\mathbf{a}^\downarrow, \mathbf{a}^\uparrow) = \text{AF}(t_1^{|\mathcal{C}|} \sqsubseteq t_2^{|\mathcal{C}|}, \text{AF}(t_1^{|\mathcal{C}|-1} \sqsubseteq t_2^{|\mathcal{C}|-1}, \dots, \text{AF}(t_1^1 \sqsubseteq t_2^1, \mathbf{a}^\downarrow, \mathbf{a}^\uparrow) \dots)).$$

The sequential composition preserves homomorphism for  $\text{AF}_\mathcal{C}$ . In Sect. 5 we tacitly assume that for arbitrary terms the function  $\text{AF}_\mathcal{C}$  is defined in a similar way.

## 4.2 Fixed-Point Algorithm

Now we present the algorithm (see Algorithm 1) that computes a chain of approximations for the case  $\mathbf{V}^b(\mathcal{C}) = \emptyset$  that converges to the solution if one exists.

**Theorem 1 (Termination).** *For any set of constraints  $\mathcal{C}$  such that  $\mathbf{V}^b(\mathcal{C}) = \emptyset$ , Algorithm 1 terminates after a finite number of steps.*

*Proof.*  $\text{AF}_\mathcal{C}$  is a monotonic function that maps  $\mathbf{a}_{i-1}^\downarrow \in \tilde{\mathcal{T}}_m^\downarrow$  to  $\mathbf{a}_i^\downarrow \in \tilde{\mathcal{T}}_m^\downarrow$  and  $\mathbf{a}_{i-1}^\uparrow \in \tilde{\mathcal{T}}_n^\uparrow$  to  $\mathbf{a}_i^\uparrow \in \tilde{\mathcal{T}}_n^\uparrow$ , where  $\mathbf{a}_{i-1}^\downarrow$  and  $\mathbf{a}_i^\downarrow$  are elements of the lattice  $(\tilde{\mathcal{T}}_m^\downarrow, \sqsubseteq)$  such that  $\mathbf{a}_i^\downarrow \sqsubseteq \mathbf{a}_{i-1}^\downarrow$ , and  $\mathbf{a}_i^\uparrow$  and  $\mathbf{a}_{i-1}^\uparrow$  are elements of the lattice  $(\tilde{\mathcal{T}}_n^\uparrow, \sqsubseteq)$  such that  $\mathbf{a}_{i-1}^\uparrow \sqsubseteq \mathbf{a}_i^\uparrow$ . Therefore, Algorithm 1, which iteratively calls  $\text{AF}_\mathcal{C}$ , terminates after a finite number of steps if the lattices have a finite height.

The semilattice for symbols has a fixed height of two element (the nil and the symbol itself). The rest of the terms, which represent collections, may “expand” only a finite number of times for a given  $\mathcal{C}$  (by expansion we mean adding new elements to a collection, which leads to term coercion in the semilattice). The size of a tuple is fixed. A record and a choice are expanded by adding elements

---

**Algorithm 1** CSP-WS( $\mathcal{C}$ ), where  $V^b(\mathcal{C}) = \emptyset$

---

```

1:  $i \leftarrow 0$ 
2:  $(\mathbf{a}_0^\downarrow, \mathbf{a}_0^\uparrow) \leftarrow ((\text{nil}, \dots, \text{nil}), (\text{none}, \dots, \text{none}))$ 
3: repeat
4:    $i \leftarrow i + 1$ 
5:    $(\mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow) \leftarrow \text{AF}_C(\mathbf{a}_{i-1}^\downarrow, \mathbf{a}_{i-1}^\uparrow)$ 
6: until  $(\mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow) = (\mathbf{a}_{i-1}^\downarrow, \mathbf{a}_{i-1}^\uparrow)$ 
7: if  $(\mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow) = (\perp, \top)$  then
8:   return Unsat
9: else
10:  return  $(\mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow)$ 
11: end if

```

---

with labels that are not yet presented in the collection. The set of labels in  $\mathcal{C}$  is finite and the algorithm cannot generate new labels. Therefore, the record and the choice can expand only a finite number of times. Therefore, the lattices have a finite height.

Substitution of variables with ground terms is a monotonic function. Below we prove that the substitution of down-coerced variables is a decreasing function. Similarly, we can prove that the substitution of up-coerced variables is an increasing function in the same way.

**Proposition 2 (Substitution monotonicity).** *Let  $t$  be a term such that  $|V^b(t)| = \emptyset$ ,  $\mathbf{v}^\downarrow = (v_1, \dots, v_k)$  be a vector of down-coerced variables in  $t$ , and  $\mathbf{s}_1^\downarrow = (s_1^1, \dots, s_k^1)$  and  $\mathbf{s}_2^\downarrow = (s_1^2, \dots, s_k^2)$  be vectors of down-coerced ground terms such that  $\mathbf{s}_1^\downarrow \sqsubseteq \mathbf{s}_2^\downarrow$ . Then*

$$t[\mathbf{v}^\downarrow / \mathbf{s}_1^\downarrow] \sqsubseteq t[\mathbf{v}^\downarrow / \mathbf{s}_2^\downarrow].$$

*Proof.* The monotonicity of the substitution follows from the structure of the seniority relation. Any term is covariant with respect to its subterms (see Definition 3).

The function AF produces the “tightest” approximation. Furthermore, the tightest approximation is unique.

**Lemma 2.** *Assume a constraint  $t_1 \sqsubseteq t_2$  and approximations  $(\mathbf{a}_1^\downarrow, \mathbf{a}_1^\uparrow)$  and  $(\mathbf{a}_2^\downarrow, \mathbf{a}_2^\uparrow)$  such that  $\text{AF}(t_1 \sqsubseteq t_2, \mathbf{a}_1^\downarrow, \mathbf{a}_1^\uparrow) = (\mathbf{a}_2^\downarrow, \mathbf{a}_2^\uparrow)$  are given. If*

$$t_1[\mathbf{v}^\downarrow / \mathbf{a}_2^\downarrow, \mathbf{v}^\uparrow / \mathbf{a}_1^\uparrow] \sqsubseteq t_2[\mathbf{v}^\downarrow / \mathbf{a}_1^\downarrow, \mathbf{v}^\uparrow / \mathbf{a}_2^\uparrow],$$

*then:*

1. *No approximation  $(\mathbf{a}_3^\downarrow, \mathbf{a}_3^\uparrow)$  exists such that  $(\mathbf{a}_3^\downarrow, \mathbf{a}_3^\uparrow) \neq (\mathbf{a}_2^\downarrow, \mathbf{a}_2^\uparrow)$ ,  $\mathbf{a}_2^\downarrow \sqsubseteq \mathbf{a}_3^\downarrow$ ,  $\mathbf{a}_3^\uparrow \sqsubseteq \mathbf{a}_2^\uparrow$  and*

$$t_1[\mathbf{v}^\downarrow / \mathbf{a}_2^\downarrow, \mathbf{v}^\uparrow / \mathbf{a}_1^\uparrow] \sqsubseteq t_1[\mathbf{v}^\downarrow / \mathbf{a}_3^\downarrow, \mathbf{v}^\uparrow / \mathbf{a}_1^\uparrow] \sqsubseteq t_2[\mathbf{v}^\downarrow / \mathbf{a}_1^\downarrow, \mathbf{v}^\uparrow / \mathbf{a}_3^\uparrow]. \quad (1)$$

2. For any other approximation  $(\mathbf{a}'_2^\downarrow, \mathbf{a}'_2^\uparrow)$  such that

$$t_1[\mathbf{v}^\downarrow/\mathbf{a}'_2^\downarrow, \mathbf{v}^\uparrow/\mathbf{a}'_1^\uparrow] \sqsubseteq t_2[\mathbf{v}^\downarrow/\mathbf{a}'_1^\downarrow, \mathbf{v}^\uparrow/\mathbf{a}'_2^\uparrow] \quad (2)$$

there exists  $(\mathbf{a}'_3^\downarrow, \mathbf{a}'_3^\uparrow)$  such that

$$t_1[\mathbf{v}^\downarrow/\mathbf{a}'_2^\downarrow, \mathbf{v}^\uparrow/\mathbf{a}'_1^\uparrow] \sqsubseteq t_1[\mathbf{v}^\downarrow/\mathbf{a}'_3^\downarrow, \mathbf{v}^\uparrow/\mathbf{a}'_1^\uparrow] \sqsubseteq t_2[\mathbf{v}^\downarrow/\mathbf{a}'_1^\downarrow, \mathbf{v}^\uparrow/\mathbf{a}'_3^\uparrow].$$

*Proof.* 1. By the definition in Sect. 4.1 the function  $\text{AF}$  coerces the approximation  $(\mathbf{a}_1^\downarrow, \mathbf{a}_1^\uparrow)$  only if the coercion is required satisfaction of  $t_1 \sqsubseteq t_2$ . The function produces  $(\mathbf{a}_2^\downarrow, \mathbf{a}_2^\uparrow)$  as a result. The approximation  $(\mathbf{a}_3^\downarrow, \mathbf{a}_3^\uparrow)$  such that (1) holds could only exist if  $\text{AF}$  performed excessive coercions, which always can be avoided.

2. The uniqueness of  $(\mathbf{a}_2^\downarrow, \mathbf{a}_2^\uparrow)$  follows from the definition of the seniority relation (Definition 3).

**Lemma 3.** Assume a set of constraints  $\mathcal{C}$ ,  $\mathbb{V}^b(\mathcal{C}) = \emptyset$ , is given. Let for  $k > 0$

$$(\mathbf{a}_0^\downarrow, \mathbf{a}_0^\uparrow), \dots, (\mathbf{a}_k^\downarrow, \mathbf{a}_k^\uparrow)$$

be a chain of approximations such that  $(\mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow) = \text{AF}_\mathcal{C}(\mathbf{a}_{i-1}^\downarrow, \mathbf{a}_{i-1}^\uparrow)$  for any  $0 < i \leq k$ , and  $\mathbf{a}_0^\downarrow = (\text{nil}, \dots, \text{nil})$  and  $\mathbf{a}_0^\uparrow = (\text{none}, \dots, \text{none})$ . Then for any fixed-point  $(\mathbf{a}^\downarrow, \mathbf{a}^\uparrow)$

$$\mathbf{a}^\downarrow \sqsubseteq \mathbf{a}_k^\downarrow \text{ and } \mathbf{a}^\uparrow \sqsubseteq \mathbf{a}_k^\uparrow. \quad (3)$$

*Proof.* The proof consists of two parts. First, we prove that a fixed-point  $(\mathbf{s}^\downarrow, \mathbf{s}^\uparrow)$  with property (3) exists. Then we show that the  $\text{AF}_\mathcal{C}$  converges to  $(\mathbf{s}^\downarrow, \mathbf{s}^\uparrow)$ , i.e.  $(\mathbf{a}_k^\downarrow, \mathbf{a}_k^\uparrow) = (\mathbf{s}^\downarrow, \mathbf{s}^\uparrow)$ .

**Existence**  $(\tilde{\mathcal{T}}_m^\downarrow, \sqsubseteq)$  and  $(\tilde{\mathcal{T}}_n^\uparrow, \sqsubseteq)$  are complete lattices and  $\text{AF}_\mathcal{C}$  is an order-preserving function. By Knaster-Tarski theorem [20], the sets of fixed points of  $\text{AF}_\mathcal{C}$  in  $(\tilde{\mathcal{T}}_m^\downarrow, \sqsubseteq)$  and  $(\tilde{\mathcal{T}}_n^\uparrow, \sqsubseteq)$  are complete lattices too. Therefore, there exists the fixed-point  $(\mathbf{s}^\downarrow, \mathbf{s}^\uparrow)$  such that for any fixed-point  $(\bar{\mathbf{s}}^\downarrow, \bar{\mathbf{s}}^\uparrow)$ ,  $\bar{\mathbf{s}}^\downarrow \sqsubseteq \mathbf{s}^\downarrow$  and  $\mathbf{s}^\downarrow \sqsubseteq \bar{\mathbf{s}}^\downarrow$ .

**Reachability** Proof by contradiction. Assume that  $\text{AF}_\mathcal{C}$  does not converge to  $(\mathbf{s}^\downarrow, \mathbf{s}^\uparrow)$ , i.e.  $(\mathbf{a}_k^\downarrow, \mathbf{a}_k^\uparrow) = (\bar{\mathbf{s}}^\downarrow, \bar{\mathbf{s}}^\uparrow)$ , where  $(\bar{\mathbf{s}}^\downarrow, \bar{\mathbf{s}}^\uparrow) \neq (\mathbf{s}^\downarrow, \mathbf{s}^\uparrow)$ , and  $\bar{\mathbf{s}}^\downarrow \sqsubseteq \mathbf{s}^\downarrow$  or  $\mathbf{s}^\downarrow \sqsubseteq \bar{\mathbf{s}}^\downarrow$ . Let  $\bar{\mathbf{s}}^\downarrow \sqsubseteq \mathbf{s}^\downarrow$  (the case when  $\mathbf{s}^\downarrow \sqsubseteq \bar{\mathbf{s}}^\downarrow$  is considered similarly).

Let  $(\mathbf{a}_{i-1}^\downarrow, \mathbf{a}_{i-1}^\uparrow)$  be the approximation that precedes  $(\bar{\mathbf{s}}^\downarrow, \bar{\mathbf{s}}^\uparrow)$  in the chain of approximations:  $\text{AF}_\mathcal{C}(\mathbf{a}_{i-1}^\downarrow, \mathbf{a}_{i-1}^\uparrow) = (\bar{\mathbf{s}}^\downarrow, \bar{\mathbf{s}}^\uparrow)$ . For every constraint  $t_1 \sqsubseteq t_2 \in \mathcal{C}$

$$t_1[\mathbf{v}^\downarrow/\bar{\mathbf{s}}^\downarrow, \mathbf{v}^\uparrow/\bar{\mathbf{a}}_{i-1}^\uparrow] \sqsubseteq t_2[\mathbf{v}^\downarrow/\mathbf{a}_{i-1}^\downarrow, \mathbf{v}^\uparrow/\bar{\mathbf{s}}^\uparrow].$$

Since  $\mathbf{s}^\downarrow$  is a fixed point, then

$$t_1[\mathbf{v}^\downarrow/\mathbf{s}^\downarrow, \mathbf{v}^\uparrow/\mathbf{a}_{i-1}^\uparrow] \sqsubseteq t_2[\mathbf{v}^\downarrow/\mathbf{a}_{i-1}^\downarrow, \mathbf{v}^\uparrow/\mathbf{s}^\uparrow].$$

On the other hand,  $\bar{\mathbf{s}}^\downarrow \sqsubseteq \mathbf{s}^\downarrow$ . Due to substitution monotonicity (Proposition 2),

$$t_1[\mathbf{v}^\downarrow/\bar{\mathbf{s}}^\downarrow, \mathbf{v}^\uparrow/\mathbf{a}_{i-1}^\uparrow] \sqsubseteq t_1[\mathbf{v}^\downarrow/\mathbf{s}^\downarrow, \mathbf{v}^\uparrow/\mathbf{a}_{i-1}^\uparrow]. \quad (4)$$

It contradicts Lemma 2, which states that  $\text{AF}_C$  produces the “tightest” approximation  $\bar{s}^\downarrow$ , but according to (4) it follows that  $s^\downarrow$  is the “tightest”.

Therefore,  $\text{AF}_C$  converges to the fixed point  $s^\downarrow$  and no  $\bar{s}^\downarrow$  exists such that  $s^\downarrow \subseteq \bar{s}^\downarrow$ .

**Theorem 2 (Correctness).** *For any set of constraints  $C$  such that  $V^b(C) = \emptyset$ , CSP-WS for  $C$  is unsatisfiable iff Algorithm 1 returns **Unsat**.*

*Proof.* Proof by contradiction.

( $\Rightarrow$ ) Let  $C$  be an unsatisfiable set of constraints and Algorithm 1 returns  $(s^\downarrow, s^\uparrow)$  such that  $(s^\downarrow, s^\uparrow) \neq (\underline{\perp}, \overline{\top})$ .  $(s^\downarrow, s^\uparrow)$  is the fixed point that contains values satisfying  $C$ . This contradicts the initial hypothesis. Therefore, Algorithm 1 returns **Unsat** if  $C$  is unsatisfiable.

( $\Leftarrow$ ) Let Algorithm 1 return **Unsat** and  $C$  has a solution. In this case the chain of approximations in Algorithm 1 returns  $(\underline{\perp}, \overline{\top})$ . This is the fixed point and by Lemma 3 no other fixed point  $(\bar{s}^\downarrow, \bar{s}^\uparrow)$  exists such that  $\underline{\perp} \subseteq \bar{s}^\downarrow$  or  $\bar{s}^\uparrow \subseteq \overline{\top}$ , which means that no fixed points apart from  $(\underline{\perp}, \overline{\top})$  exists. This contradicts the initial hypothesis. Therefore,  $C$  is unsatisfiable if Algorithm 1 returns **Unsat**.

## 5 CSP-WS Algorithm

A straightforward algorithm for CSP-WS has to run Algorithm 1 for each of  $2^l$  pairs of the semi-lattices, where  $l = |V^b(C)|$ . Instead, we present iterative Algorithm 2 which takes the advantage of the order-theoretical structure of the MDL and generates an adjunct SAT problem on the way.

---

### Algorithm 2 CSP-WS( $C$ )

---

```

1:  $c \leftarrow |C|$ 
2:  $i \leftarrow 0$ 
3:  $B_0 \leftarrow \emptyset$ 
4:  $\mathbf{a}_0^\downarrow \leftarrow (\text{nil}, \dots, \text{nil})$ 
5:  $\mathbf{a}_0^\uparrow \leftarrow (\text{none}, \dots, \text{none})$ 
6: repeat
7:    $i \leftarrow i + 1$ 
8:    $(\mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow) \leftarrow \text{AF}_C(\mathbf{a}_{i-1}^\downarrow, \mathbf{a}_{i-1}^\uparrow)$ 
9:    $B_i \leftarrow B_{i-1} \cup \bigcup_{t_1 \sqsubseteq t_2 \in C} (\text{WFC}(t_1[v/\mathbf{a}_i]) \cup \text{WFC}(t_2[v/\mathbf{a}_i]) \cup \text{SC}(t_1[v/\mathbf{a}_i] \sqsubseteq t_2[v/\mathbf{a}_i]))$ 
10: until  $(\text{SAT}(B_i), \mathbf{a}_i^\downarrow, \mathbf{a}_i^\uparrow) = (\text{SAT}(B_{i-1}), \mathbf{a}_{i-1}^\downarrow, \mathbf{a}_{i-1}^\uparrow)$ 
11: if  $B_i$  is unsatisfiable then
12:   return Unsat
13: else
14:   return  $(b, \mathbf{a}_i^\downarrow[f/b], \mathbf{a}_i^\uparrow[f/b])$ , where  $b \in \text{SAT}(B_i)$ 
15: end if
```

---

Let  $B_0 \subseteq B_1 \subseteq \dots \subseteq B_s$  be sets of Boolean constraints, and  $\mathbf{a}^\downarrow$  and  $\mathbf{a}^\uparrow$  be vectors of semiground terms such that  $|\mathbf{a}^\downarrow| = |V^\downarrow(C)|$  and  $|\mathbf{a}^\uparrow| = |V^\uparrow(C)|$ . We

1.  $\text{WFC}(t) = \emptyset$  if  $t$  is a symbol;
2.  $\text{WFC}(t) = \bigcup_{1 \leq i \leq n} \text{WFC}(t_i)$  if  $t$  is a tuple  $(t_1 \dots t_n)$ ;
3.  $\text{WFC}(t) = \{\neg(g_i \wedge g_j) \mid 1 \leq i \neq j \leq n \text{ and } l_i = l_j\} \cup \bigcup_{1 \leq i \leq n} \{g_i \rightarrow g \mid g \in \text{WFC}(t_i)\}$  if  $t$  is a record  $\{l_1(g_1): t_1, \dots, l_n(g_n): t_n\}$  or a choice  $(:l_1(g_1): t_1, \dots, l_n(g_n): t_n:)$ ;
4.  $\text{WFC}(t) = \{\neg(g_i \wedge g_j) \mid 1 \leq i \neq j \leq n\} \cup \{\bigvee_{1 \leq i \leq n} g_i\} \cup \bigcup_{1 \leq i \leq n} \{g_i \rightarrow g \mid g \in \text{WFC}(t_i)\}$  if  $t$  is a switch  $\langle (g_1): t_1, \dots, (g_n): t_n \rangle$ .

**Figure 3.** The set of Boolean constraints that ensures well-formedness of a term  $t$

1.  $\text{SC}(t_1 \sqsubseteq t_2) = \emptyset$ , if  $t_1$  and  $t_2$  are equal symbols.
2.  $\text{SC}(t_1 \sqsubseteq t_2) = \bigcup_{1 \leq i \leq k} \text{SC}(t_i^1 \sqsubseteq t_i^2)$ , if  $t_1$  is a tuple  $(t_1^1 \dots t_k^1)$  and  $t_2$  is a tuple  $(t_1^2 \dots t_k^2)$ ;
3.  $\text{SC}(t_1 \sqsubseteq t_2) = \bigcup_{1 \leq j \leq m} \text{SC}_j(t_j^2)$ , if  $t_1$  is a record  $\{l_1^1(g_1^1): t_1^1, \dots, l_k^1(g_k^1): t_k^1\}$ ,  $t_2$  is a record  $\{l_1^2(g_1^2): t_1^2, \dots, l_m^2(g_m^2): t_m^2\}$  and  $\text{SC}_j(t_j^2)$  is one of the following:
  - (a)  $\text{SC}_j(t_j^2) = \{(g_i^1 \wedge g_j^2) \rightarrow g \mid g \in \text{SC}(t_i^1 \sqsubseteq t_j^2)\}$ , if  $\exists i: 1 \leq i \leq k$  and  $l_i^1 = l_j^2$ ;
  - (b)  $\text{SC}_j(t_j^2) = \{\neg g_j^2\}$ , otherwise;
4.  $\text{SC}(t_1 \sqsubseteq t_2) = \bigcup_{1 \leq i \leq m} \text{SC}_i(t_i^1)$ , if  $t_1$  is a choice  $(:l_1^1(g_1^1): t_1^1, \dots, l_k^1(g_k^1): t_k^1:)$ ,  $t_2$  is a choice  $(:l_1^2(g_1^2): t_1^2, \dots, l_m^2(g_m^2): t_m^2:)$  and  $\text{SC}_i(t_i^1)$  is one of the following:
  - (a)  $\text{SC}_i(t_i^1) = \{(g_i^1 \wedge g_j^2) \rightarrow g \mid g \in \text{SC}(t_i^1 \sqsubseteq t_j^2)\}$ , if  $\exists j: 1 \leq j \leq m$  and  $l_i^1 = l_j^2$ ;
  - (b)  $\text{SC}_i(t_i^1) = \{\neg g_i^1\}$ , otherwise;
5.  $\text{SC}(t_1 \sqsubseteq t_2) = \{g_i^1 \rightarrow g \mid 1 \leq i \leq k \text{ and } g \in \text{SC}(t_i^1 \sqsubseteq t_i^2)\}$ , if  $t_1$  is a switch  $\langle (g_1^1): t_1^1, \dots, (g_k^1): t_k^1 \rangle$  and  $t_2$  is an arbitrary term.
6.  $\text{SC}(t_1 \sqsubseteq t_2) = \{g_i^2 \rightarrow g \mid 1 \leq i \leq k \text{ and } g \in \text{SC}(t_1 \sqsubseteq t_i^2)\}$ , if  $t_1$  is an arbitrary term and  $t_2$  is a switch  $\langle (g_1^2): t_1^2, \dots, (g_k^2): t_k^2 \rangle$ .
7.  $\text{SC}(t_1 \sqsubseteq t_2) = \{\text{false}\}$ , otherwise.

**Figure 4.** The set of Boolean constraints that ensures the seniority relation  $t_1 \sqsubseteq t_2$

seek the solution as a fixed point of a chain of approximations in the following form:

$$(\mathbf{B}_0, \mathbf{a}_0^\downarrow, \mathbf{a}_0^\uparrow), \dots, (\mathbf{B}_{s-1}, \mathbf{a}_{s-1}^\downarrow, \mathbf{a}_{s-1}^\uparrow), (\mathbf{B}_s, \mathbf{a}_s^\downarrow, \mathbf{a}_s^\uparrow),$$

where for every  $1 \leq i \leq s$  and a vector of Boolean values  $\mathbf{b}$  that is a solution to  $\text{SAT}(\mathbf{B}_i)$ :

$$\mathbf{a}_i^\downarrow[\mathbf{f}/\mathbf{b}] \sqsubseteq \mathbf{a}_{i-1}^\downarrow[\mathbf{f}/\mathbf{b}] \quad \text{and} \quad \mathbf{a}_{i-1}^\uparrow[\mathbf{f}/\mathbf{b}] \sqsubseteq \mathbf{a}_i^\uparrow[\mathbf{f}/\mathbf{b}].$$

The adjunct set of Boolean constraints potentially expands at every iteration of the algorithm by inclusion of further logic formulas produced by the set of Boolean constraint  $\text{WFC}$  (see Fig. 3) ensuring well-formedness of the terms and the set of Boolean constraints  $\text{SC}$  (see Fig. 4) ensuring that the seniority relations holds. The starting point is  $\mathbf{B}_0 = \emptyset$ ,  $\mathbf{a}_0^\downarrow = (\text{nil}, \dots, \text{nil})$ ,  $\mathbf{a}_0^\uparrow = (\text{none}, \dots, \text{none})$  and the chain terminates as soon as  $\text{SAT}(\mathbf{B}_s) = \text{SAT}(\mathbf{B}_{s-1})$ ,  $\mathbf{a}_s^\uparrow = \mathbf{a}_{s-1}^\uparrow$ ,  $\mathbf{a}_s^\downarrow = \mathbf{a}_{s-1}^\downarrow$ , where by  $\text{SAT}(\mathbf{B}_i)$  we mean a set of Boolean vector satisfying  $\mathbf{B}_i$ . Whether the set of Boolean constraints actually expands or not can be determined by checking the satisfiability of  $\text{SAT}(\mathbf{B}_i) \neq \text{SAT}(\mathbf{B}_{i-1})$  for the current iteration  $i$ .



We argue that if the original CSP-WS is satisfiable then so is  $\text{SAT}(\mathbf{B}_s)$  and that the tuple of vectors  $(\mathbf{b}_s, \mathbf{a}_s^\downarrow[\mathbf{f}/\mathbf{b}_s], \mathbf{a}_s^\uparrow[\mathbf{f}/\mathbf{b}_s])$  is a solution to the former, where  $\mathbf{b}_s$  is a solution of  $\text{SAT}(\mathbf{B}_s)$ . In other words, the iterations terminate when the conditional approximation limits the t-variables, and when the adjunct SAT constrains the b-variables enough to ensure the satisfaction of all CSP-WS constraints. In general, the set  $\text{SAT}(\mathbf{B}_s)$  can have more than one solution and we select one of them. Heuristics that allows to choose a solution that is better for the given application is left for further research.

**Implementation.** We implemented the CSP-WS algorithm as a solver in the OCaml language. The input for the solver is a set of constraints and the output is in the form of assignments to b-variables and t-variables. It works on top of the PicoSAT [3] library (although any other SAT solver could be used instead). PicoSAT is employed as a subsolver that deals with Boolean assertions.

## 6 Conclusion and Future Work

We have presented a new mechanism for choreographing service interfaces based on CSP and SAT that configures generic non-local interfaces in the context. We developed a Message Definition Language that can be used in the context of service-based applications. Our mechanism supports subtyping, polymorphism and inheritance, thanks to the order relation defined on MDL terms. We presented the CSP solution algorithm for interface configuration, which has been developed specifically for this problem.

In the context of Cloud, our results may prove useful to the software-as-a-service community since we can support much more generic interfaces than are currently available. Building services the way we do could enable service providers to configure a solution for a network customer based on services that they have at their disposal as well as those provided by other providers and the customer themselves, all solely on the basis of interface definitions and automatic tuning to non-local requirements.

The next step will be the design of a mechanism for automatic interface derivation from code of the services, which can be done in a straightforward manner. This brings an advantage over choreography mechanisms that rely on behavioural protocols: automatic derivation of the behaviour from the code is a difficult problem that have not been solved yet.

## References

1. Arkin, A., Askary, S., Fordin, S., Jekeli, W., Kawaguchi, K., Orchard, D., Pogliani, S., Riemer, K., Struble, S., Takacs-Nagy, P., et al.: Web service choreography interface (WSCl) 1.0. Standards proposal by BEA Systems, Intalio, SAP, and Sun Microsystems (2002)
2. Barros, A., Dumas, M., Oaks, P.: A critical overview of the web services choreography description language. BPTrends Newsletter 3, 1–24 (2005)
3. Biere, A.: Picosat essentials. Journal on Satisfiability, Boolean Modeling and Computation (JSAT) (2008)

4. Bouguettaya, A., Sheng, Q.Z., Daniel, F.: Advanced web services. Springer (2013)
5. Bourne, S., Szabo, C., Sheng, Q.Z.: Ensuring well-formed conversations between control and operational behaviors of web services. In: Service-Oriented Computing, pp. 507–515. Springer (2012)
6. Carbone, M., Honda, K., Yoshida, N.: Structured communication-centred programming for web services. In: Programming Languages and Systems, pp. 2–17. Springer (2007)
7. Coppo, M., Dezani-Ciancaglini, M., Padovani, L., Yoshida, N.: A gentle introduction to multiparty asynchronous session types. In: Formal Methods for Multicore Programming, pp. 146–178. Springer (2015)
8. Davies, R., Pfenning, F.: Intersection types and computational effects. In: ACM Sigplan Notices. vol. 35, pp. 198–208. ACM (2000)
9. Duan, Q., Yan, Y., Vasilakos, A.V.: A survey on service-oriented network virtualization toward convergence of networking and cloud computing. Network and Service Management, IEEE Transactions on 9(4), 373–392 (2012)
10. Dustdar, S., Schreiner, W.: A survey on web services composition. International journal of web and grid services 1(1), 1–30 (2005)
11. Gaster, B.R., Jones, M.P.: A polymorphic type system for extensible records and variants (1996)
12. Grelck, C., Scholz, S.B., Shafarenko, A.: A gentle introduction to s-net: Typed stream processing and declarative coordination of asynchronous components. Parallel Processing Letters 18(02), 221–237 (2008)
13. Group, W.S.C.W., et al.: Web services choreography description language (2002)
14. Honda, K., Yoshida, N., Carbone, M.: Multiparty asynchronous session types. ACM SIGPLAN Notices 43(1), 273–284 (2008)
15. Kildall, G.A.: A unified approach to global program optimization. In: Proceedings of the 1st annual ACM SIGACT-SIGPLAN symposium on Principles of programming languages. pp. 194–206. ACM (1973)
16. Leijen, D.: Extensible records with scoped labels. Trends in Functional Programming 5, 297–312 (2005)
17. Mathew, S.S., Atif, Y., Sheng, Q.Z., Maamar, Z.: The web of things-challenges and enabling technologies. In: Internet of things and inter-cooperative computational technologies for collective intelligence, pp. 1–23. Springer (2013)
18. Sheng, Q.Z., Maamar, Z., Yao, L., Szabo, C., Bourne, S.: Behavior modeling and automated verification of web services. Information Sciences 258, 416–433 (2014)
19. Sheng, Q.Z., Qiao, X., Vasilakos, A.V., Szabo, C., Bourne, S., Xu, X.: Web services composition: A decades overview. Information Sciences 280, 218–238 (2014)
20. Tarski, A., et al.: A lattice-theoretical fixpoint theorem and its applications. Pacific journal of Mathematics 5(2), 285–309 (1955)
21. Zheng, Z., Lyu, M.R.: Personalized reliability prediction of web services. ACM Transactions on Software Engineering and Methodology (TOSEM) 22(2), 12 (2013)